Évaluation intrinsèque et extrinsèque du nettoyage de pages Web

Gaël Lejeune, Romain Brixtel, Charlotte Lecluze

Évaluation intrinsèque et extrinsèque du nettoyage de pages Web

Gaël Lejeune[1], Romain Brixtel [2], Charlotte Lecluze[3]
(1) LINA, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France
(2) Université de Lausanne – HEC - Département de comportement organisationnel, Quartier Dorigny, 1015 Lausanne, Suisse
(3) GREYC, Campus Côte de Nacre, Boulevard du Maréchal Juin, 14032 CAEN cedex 5, France
prenom.nom@univ-nantes.fr, unil.ch, unicaen.fr

Résumé. Le nettoyage de documents issus du web est une tâche importante pour le TAL en général et pour la constitution de corpus en particulier. Cette phase est peu traitée dans la littérature, pourtant elle n'est pas sans influence sur la qualité des informations extraites des corpus. Nous proposons deux types d'évaluation de cette tâche de *détourage* : (I) une évaluation intrinsèque fondée sur le contenu en mots, balises et caractères ; (II) une évaluation extrinsèque fondée sur la tâche, en examinant l'effet du détourage des documents sur le système placé en aval de la chaîne de traitement. Nous montrons que les résultats ne sont pas cohérents entre ces deux évaluations ainsi qu'entre les différentes langues. Ainsi, le choix d'un outil de détourage devrait être guidé par la tâche visée plutôt que par la simple évaluation intrinsèque.

Abstract.
Intrinsic and extrinsic evaluation of boilerplate removal tools
In this article, we tackle the problem of evaluation of web page cleaning tools. This task is seldom studied in the literature although it has consequences on the linguistic processing performed on web-based corpora. We propose two types of evaluation : (I) an intrinsic (content-based) evaluation with measures on words, tags and characters ; (II) an extrinsic (task-based) evaluation on the same corpus by studying the effects of the cleaning step on the performances of an NLP pipeline. We show that the results are not consistent in both evaluations. We show as well that there are important differences in the results between the studied languages. We conclude that the choice of a web page cleaning tool should be made in view of the aimed task rather than on the performances of the tools in an intrinsic evaluation.

Mots-clés : Nettoyage de pages Web, collecte de corpus, évaluation intrinsèque, évaluation extrinsèque, détourage.

Keywords: Web page cleaning, corpus collecting, intrinsic evaluation, extrinsic evaluation, web scraping.

1 Introduction

La quantité grandissante de documents numériques disponibles permet de disposer de corpus pour différentes tâches de Traitement Automatique des Langues (TAL). Cependant, le format des documents n'est pas toujours adapté aux modules placés en aval de la chaîne de traitement de TAL. Ces documents contiennent des éléments non-informatifs qu'il convient de détecter pour faciliter les analyses ultérieures. Aussi, un même rendu peut être généré par des codes sources différents : il n'y a pas de biunivocité source-rendu. L'opération d'extraction du contenu des documents HTML peut être nommée suppression du squelette de page (*boilerplate removal*), détection de modèle (*Web Page Template Detection*) ou plus généralement nettoyage (*Web Page Cleaning*) de page Web. Toutefois ces termes, et en particulier « nettoyage », sont réducteurs vis-à-vis de l'importance de cette étape dans la chaîne de traitement. Nous proposons d'utiliser le terme de *détourage*. Issu de la photographie, ce terme désigne le fait de n'extraire d'une illustration que les parties utiles. Le détourage de pages Web consiste à extraire le texte recherché à partir des données brutes et du rendu tout en conservant certaines données de structure (titraison, paragraphes…). Cette opération est effectuée par un *détoureur*. Nous proposons dans cet article deux types d'évaluation pour le détourage :

Évaluation intrinsèque Nous utilisons une évaluation fondée sur le contenu détouré. Cette modalité d'évaluation est la plus fréquente dans la littérature (Endrédy & Novák, 2013). Nous exploitons les métriques de la compétition CLEANEVAL (Baroni *et al.*, 2008) : distance d'édition au niveau des mots avec ou sans balise(s). Nous y ajoutons une distance d'édition sur les caractères pour permettre une meilleure évaluation sur le chinois.

Évaluation extrinsèque Nous proposons une évaluation par la tâche qui consiste à mesurer la qualité d'un détoureur en fonction des résultats obtenus par un système placé en aval de la chaîne de traitement. Nous utilisons pour ce faire DANIEL (Brixtel *et al.*, 2013), un système de veille multilingue, ainsi que le corpus de référence associé.

Le corpus que nous utilisons est composé d'articles de presse uniquement, cible principale de la campagne CLEANEVAL. Dans la Section 2, nous exposons la problématique du détourage. Puis nous détaillons dans la Section 3 les caractéristiques des différents détoureurs comparés. La Section 4 est consacrée à la description du corpus et à l'évaluation. La Section 5 propose une conclusion de notre travail et des perspectives d'évolution.

2 La problématique du détourage des pages Web

Pour l'humain, la détection du contenu purement textuel ne semble pas poser de difficulté. Le contenu apparaît la plupart du temps au centre de la page et le titre permet de fixer rapidement un point de départ pour la lecture. Toutefois, automatiser ce processus de reconnaissance du contenu textuel reste un défi à l'heure actuelle. Ceci est illustré par la Figure 1 qui présente deux exemples de détourages erronés sur deux sources différentes dans les premiers résultats de *Google News* [1].

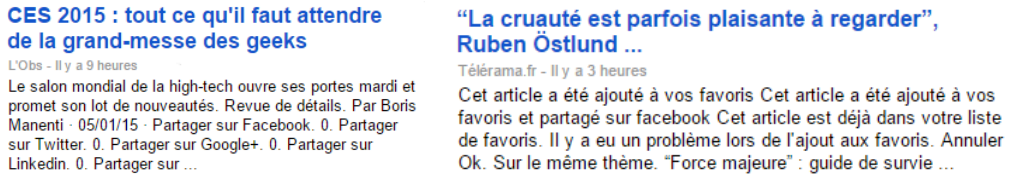

FIGURE 1: Deux exemples d'erreurs de détourage sur des chapeaux d'articles tirés de *Google News*.

Le processus de détourage peut être décomposé en deux sous-tâches : le **nettoyage** (suppression du code JAVASCRIPT, du style, des menus, entêtes et pieds de page) et la **structuration** (récupération des titres, des paragraphes, des listes...). L'approche la plus intuitive pour détourer les pages Web est l'exploitation du *Document Object Model* (DOM), tel que dans les travaux de Chakrabarti *et al.* (2008) et Vieira *et al.* (2006). Dans ces derniers, différentes pages issues du même site Web sont exploitées pour identifier les similarités du DOM. Ce qui est structurellement commun correspond au contenu non-informatif (publicités, liens de navigation) et ce qui est structurellement différent constitue le contenu informatif. D'autres approches représentent la spécificité des pages Web d'un même site sous la forme d'un arbre. La position relative des nœuds permet alors de classer les différents segments dans la catégorie « informatif » ou « non-informatif » (Das *et al.*, 2012). La densité en balises HTML est une autre façon d'exploiter le code source pour classer les segments (Ferraresi *et al.*, 2008). D'autres approches exploitent la distribution des n-grammes de caractères comme dans NCLEANER (Evert, 2008) ou VICTOR (Spousta *et al.*, 2008). L'utilisation combinée de la densité en balises et de la distribution de n-grammes de caractères a été proposée par Pasternack & Roth (2009).

3 Caractéristiques des détoureurs utilisés

Les détoureurs font appel à quatre niveaux d'analyse : le site Web (caractéristiques inhérentes à différentes pages d'un même site), le rendu (simulation du rendu de la page donné par un navigateur), la structure HTML (informations de hiérarchie entre les blocs) ou le contenu textuel (phrases, mots, caractères...). Nous nous concentrons ici sur trois outils librement disponibles : `Boilerpipe`, `NCleaner` et `Justext`.

3.1 `Boilerpipe`

`Boilerpipe` [2] (Kohlschütter *et al.*, 2010) est probablement le détoureur le plus utilisé dans la communauté TAL. Il est basée sur une combinaison de critères locaux (internes aux blocs) et contextuels (relatifs aux blocs voisins). Les balises considérées comme les plus communes dans les zones textuelles sont utilisées (balises de titres `<h1>` à `<h6>`, de paragraphes `<p>` et de conteneurs `<div>`). Les balises de liens `<a>`, les mots capitalisés, les liens hypertextes ou

1. Consultés respectivement le 5 janvier 2015 et le 28 janvier 2015.
2. `http://code.google.com/p/boilerpipe/` (consulté le 1er juin 2015)

le caractère « | » sont des indicateurs de contenu non-informatif ; les points ou les virgules sont la marque de segments informatifs. Les indices contextuels se fondent sur une hypothèse de position relative des blocs de texte et de squelette : les blocs informatifs sont souvent consécutifs. L'étiquette « informatif » ou « non-informatif » d'un bloc est donc fortement dépendante de l'étiquette du bloc précédent. C'est une comparaison des densités en *tokens* (mots graphiques) par ligne dans l'affichage (largeur estimée à 80 caractères) qui permet de juger si l'étiquette du bloc doit changer.

3.2 `NCleaner`

Le détoureur `NCleaner` [3] (Evert, 2008), présenté lors de la compétition CLEANEVAL, utilise des modèles de langue en *n*-grammes de caractères. `NCleaner` mesure la probabilité qu'un caractère appartienne à la langue sachant les caractères qui le précèdent. `NCleaner` cherche à identifier les *n*-grammes (avec $1 \leq n \leq 3$) qui maximisent la probabilité d'appartenance d'un bloc au contenu informatif et ceci pour chaque langue. Trois configurations de base sont possibles :

Par défaut (NC) : Utilise un modèle *n*-gramme « indépendant » de la langue ;

Non-lexical (NCNL) : Transforme les lettres en `a` (`[:alpha:]` → `a`) et les chiffres en `0` (`[:digit:]` → `0`) ;

Avec entraînement (NCTx) : Se base sur un échantillon de x paires de documents (d_{brut},$d_{\text{détouré}}$) fournies au système, d_{brut} étant un document non détouré et $d_{\text{détouré}}$ étant un attendu de document détouré à partir de d_{brut}.

3.3 `Justext`

`Justext` [4] est un détoureur plus récent qui dépasse les résultats de BOILERPIPE selon les évaluations menées par son auteur (Pomikálek, 2011). La méthode utilisée comporte deux étapes. La première étape (dite *context-free*) consiste à collecter trois traits pour chaque bloc de texte : sa longueur en *tokens*, le nombre de liens qu'il contient et, optionnellement, la quantité de mots outils en faisant appel à une ressource externe qui existe pour 100 langues. Nous utilisons dans cet article deux configurations : avec et sans ressource(s) externe(s). En fonction de ces indices, chaque bloc reçoit une première étiquette :

Bad : Bloc de squelette.
Near good : Bloc probablement informatif.
Good : Bloc informatif.
Short : Bloc trop court pour être étiqueté.

La seconde étape (dite *context-sensitive*) consiste à étiqueter les blocs sans étiquette en fonction des étiquettes de leurs voisins. Un bloc de type *short* devient informatif s'il est entouré de blocs de la classe *good* ou *near-good*. Un bloc de type *near-good* est considéré comme informatif si le bloc qui le suit ou le bloc qui le précède est lui-même un bloc informatif.

4 Corpus et modalités d'évaluation

Nous présentons dans la Section 4.1 le format des données et les modalités d'évaluation intrinsèque proposés lors de CLEANEVAL (Baroni *et al.*, 2008). Le corpus d'évaluation ainsi que l'outil choisi pour l'évaluation extrinsèque sont présentés dans la Section 4.2. Enfin, nous décrivons les résultats obtenus dans la Section 4.3.

4.1 Format des textes et modalités d'évaluation de CLEANEVAL

Le format de référence a été obtenu par le travail d'annotateurs humains munis d'instructions précises [5]. La tâche consistait à enlever les traces du squelette de page, les codes HTML et JAVASCRIPT et ne conserver qu'une structure de texte simplifiée utilisant trois balises : `<h>` pour les titres, `<p>` pour les paragraphes et `<l>` pour les éléments de listes.

Le format de texte amène à évaluer deux aspects : la séparation du contenu et du squelette d'une part et la conservation de la structure du texte d'autre part. Le script d'évaluation de la campagne CLEANEVAL considère deux grains : les mots seuls (TO : *text only*) et les mots avec les balises (TM : *text and markup*). Pour le second cas, deux modalités sont proposées : *labelled* qui tient compte du type de balise et *unlabelled* qui n'en tient pas compte (la séquence `<p><p><l>` est alors équivalente à `<p><p><p>`). Pour chaque fichier, la version détourée automatiquement est comparée avec la

3. `http://webascorpus.sourceforge.net/PHITE.php?page=FILES_10_Software` (consulté le 1er juin 2015)
4. `http://nlp.fi.muni.cz/projects/justext/` (consulté le 1er juin 2015)
5. `http://cleaneval.sigwac.org.uk/annotation_guidelines.html` (consulté le 1er juin 2015)

version de référence. Chaque version est normalisée en deux étapes : remplacement des caractères de contrôle (sauts de lignes, tabulations ...) par des espaces et normalisation des espaces (les espaces consécutifs sont remplacés par un seul).

Chaque version est découpée à chaque signe de ponctuation ou espace en une séquence de *tokens*. Deux séquences de *tokens* sont comparées en utilisant l'algorithme de Ratcliff (Ratcliff & Metzener, 1988). Calculer la similarité entre elles revient à diviser le nombre de *tokens* en commun par le nombre de *tokens* dans les deux séquences. Les *tokens* en commun sont extraits de la plus longue sous-séquence commune, puis récursivement sur les *tokens* en commun autour de cette sous-séquence. L'algorithme de Ratcliff permet d'obtenir la liste des opérations permettant de passer d'une séquence à l'autre (Baroni *et al.*, 2008). Le fait que les métriques sur les mots et sur les balises soient intriquées gêne l'interprétation des résultats : un détoureur qui n'extrairait que les mots du texte sans structure aurait un score convenable dans la configuration TM. Deux détoureurs proches suivant l'évaluation TO peuvent présenter des résultats différents dans la configuration TM. L'utilisation du grain mot dans les deux modalités d'évaluation de CLEANEVAL pose problème pour des langues telles que le chinois. Nous introduisons donc une évaluation par caractère destinée à mieux évaluer les performances des détoureurs, sur le chinois notamment. Pour cette modalité d'évaluation, un *token* correspond à un caractère.

4.2 Corpus de référence et outil pour l'évaluation extrinsèque

L'outil DANIEL et son corpus associé ont été proposés par Brixtel *et al.* (2013) pour la classification de documents pertinents ou non-pertinents pour la veille épidémiologique. L'outil utilise à la fois des éléments de contenu et des éléments de structure, ce qui permet de mesurer la conservation de la structure à l'issue du nettoyage. Nous comparons les résultats obtenus sur les documents détourés manuellement (détourage supposé « idéal ») avec ceux obtenus sur des documents détourés automatiquement. Le corpus associé contient originellement plus de 2000 documents en cinq langues (anglais, chinois, grec, polonais et russe). Le corpus permet d'évaluer la variation en langue et en système d'écriture. Les fichiers n'étant pas directement disponibles, nous avons donc utilisé les *urls* fournies pour les télécharger [6]. Nous avons pu collecter 80% des documents bruts utilisés par DANIEL (Table 1) avec une répartition comparable entre les deux classes.

Langues	Anglais	Chinois	Grec	Polonais	Russe	Toutes
#documents du corpus d'origine	475	446	390	352	426	2089
#documents pertinents	31	16	26	30	41	144
#documents retrouvés	475 (100%)	405 (90,8%)	273 (70%)	274 (77,8%)	267 (62,7%)	1694 (81,1%)
#documents pertinents retrouvés	31 (100%)	16 (100%)	17 (65,4%)	27 (90%)	29 (70,7%)	120 (83,3%)

TABLE 1: Corpus DANIEL, documents présents dans le corpus d'origine et documents retrouvés.

4.3 Évaluation intrinsèque *vs.* évaluation extrinsèque

La Table 2 présente les résultats de différents détoureurs et de leurs combinaisons en utilisant les métriques de *Cleaneval* auxquelles nous avons ajouté une évaluation par caractère. Pour les Tables suivantes, TO désigne l'évaluation sur le texte seul (*Text Only*), TM l'évaluation sur le texte avec les balises (*Text and Markup*) et CAR désigne l'évaluation au grain caractère.

Les détoureurs utilisés sont également désignés par des abréviations : BP correspond à BOILERPIPE, JTA et JTS à JUSTEXT dans ses deux configurations (respectivement avec (JTA) et sans ressource externe (JTS)), NC est la configuration par défaut de NCLEANER, NCT5 et NCT25 sont les configurations utilisant respectivement 5 et 25 paires de textes pour l'apprentissage. Cinq paires ont été constituées manuellement pour chaque langue. Pour NCT5, les cinq paires sont utilisées pour chaque langue et une paire par langue pour le corpus complet (la paire où le document détouré est le plus long en caractères). Pour NCT25, les 25 paires sont utilisées pour chaque langue de même que pour le corpus complet. Les mesures utilisées sont le rappel (R), la précision (P) et la F_1-mesure (F_1). Les combinaisons de détoureurs consistent à exploiter le résultat d'un premier détoureur en entrée d'un second détoureur. Il s'agit d'analyser la complémentarité des détoureurs. La Table 2 présente les résultats de l'évaluation intrinsèque pour l'ensemble du corpus [7].

Sur l'évaluation intrinsèque JT est surclassé par BP dans les 3 catégories de *tokens* évalués. NC dans sa configuration par défaut présente des résultats intéressants en terme de précision mais au prix d'un rappel moindre. Les autres configurations de NC offrent des résultats peu significatifs.

La Table 3 présente les résultats de l'évaluation extrinsèque, pour chacune des langues ainsi que pour le corpus complet. La

6. `https://daniel.greyc.fr/public/api_daniel.php` (consulté le 1er juin 2015)
7. Les résultats des chaînages impliquant NCLEANER ont été omis car les performances sont faibles et n'apportent aucune information intéressante.

	TO			CAR			TM		
Mesures	P	R	F_1	P	R	F_1	P	R	F_1
BP	81.80	**88.89**	**85.20**	76.93	**81.12**	**78.97**	64.47	**85.42**	**73.48**
BP–JTA	85.01	80.20	82.54	76.94	63.86	69.79	73.30	58.74	65.22
BP–JTS	83.23	82.87	83.05	75.12	66.03	70.28	69.22	62.07	65.45
JTA	68.75	83.41	75.37	63.79	67.03	65.37	61.94	63.23	62.58
JTA–BP	72.54	85.86	78.64	69.43	73.28	71.31	66.76	69.34	68.02
JTS	62.68	86.30	72.62	56.93	68.63	62.23	54.24	66.57	59.78
JTS–BP	66.31	88.74	75.90	62.95	75.76	68.77	59.42	72.70	65.39
NC	**98.53**	39.38	56.27	**96.65**	23.15	37.36	**89.01**	30.82	45.78
NCNL	01.28	02.73	01.74	01.30	02.97	01.81	02.01	04.14	02.71
NCT5	60.43	23.83	34.18	53.81	16.03	24.70	48.41	19.89	28.19
NCT25	56.14	25.70	35.26	53.25	18.73	27.72	45.11	21.77	29.36

TABLE 2: Résultats de l'évaluation intrinsèque pour le corpus complet en cinq langues (Précision, Rappel et F_1-mesure exprimés en %) sur les grains *Text Only* (TO), CARactère (CAR) et *Text and Markup* (TM).

ligne « Référence » indique le résultat attendu à partir du détourage manuel [8]. Nous observons que l'évaluation extrinsèque établit une hiérarchie différente de celle issue de l'évaluation intrinsèque. JT, dans sa version avec ou sans ressource(s), est meilleur que BP en terme de F_1-mesure hormis sur le sous-corpus chinois. La meilleure option est cependant le chaînage BP-JT. Les résultats obtenus sur le corpus chinois sont identiques pour JTA et JTS, ceci est dû à l'absence de ressource externe pour cette langue. Ces résultats sont par ailleurs strictement équivalents à ceux obtenus pour les deux chaînages JTA-BP et JTS-BP. NC obtient des résultats intéressants sur l'anglais et dans une moindre mesure sur le polonais. Nous pouvons remarquer une grande variabilité des résultats selon les langues sur l'évaluation extrinsèque.

	Anglais			Chinois			Grec			Polonais			Russe			Corpus complet		
Mesures	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BP	60.00	25.71	36.00	78.95	**93.75**	**85.71**	**85.71**	35.94	50.00	76.47	48.15	59.09	76.19	55.17	64.00	**74.68**	47.58	58.13
BP–JTA	61.54	45.71	52.46	71.43	31.25	43.48	66.67	70.59	**68.57**	65.63	77.78	71.19	76.67	**79.31**	77.97	68.14	**62.10**	**64.98**
BP–JTS	**65.22**	42.86	51.72	71.43	31.25	43.48	63.16	70.59	66.67	64.71	**81.48**	**72.13**	74.19	**79.31**	76.67	67.54	**62.10**	64.71
JTA	55.17	45.71	50.00	66.67	37.50	48.00	59.09	**76.47**	66.67	59.26	59.26	59.26	82.14	**79.31**	**80.70**	64.35	59.68	61.92
JTA–BP	59.09	37.14	45.61	66.67	37.50	48.00	62.50	58.82	60.61	65.38	62.96	64.15	76.00	65.52	70.37	66.33	52.42	58.56
JTS	55.56	42.86	48.39	66.67	37.50	48.00	66.67	58.82	62.50	56.67	62.96	59.65	82.61	65.52	73.08	64.42	54.03	58.77
JTS–BP	60.87	40.00	48.28	66.67	37.50	48.00	66.67	47.06	55.17	62.96	62.96	62.96	78.26	62.07	69.23	67.02	50.81	57.80
NC	58.33	**60.00**	**59.15**	N/A	0.00	N/A	N/A	0.00	N/A	80.00	14.81	25.00	N/A	0.00	N/A	60.98	20.16	30.30
NCNL	50.00	14.29	22.22	N/A	0.00	N/A	N/A	0.00	N/A	23.81	18.52	20.83	**100**	6.90	12.90	26.09	9.68	14.12
NCT5	52.94	25.71	34.62	**83.33**	31.25	45.45	N/A	0.00	N/A	**82.35**	51.85	63.64	60.00	20.69	30.77	62.96	27.42	38.20
NCT25	50.00	25.71	33.96	**83.33**	31.25	45.45	20.00	5.88	9.09	**82.35**	51.85	63.64	61.54	27.59	38.09	62.71	29.84	40.44
Référence	**68.89**	**88.57**	**77.50**	**80.00**	**100**	**88.89**	**68.42**	**76.47**	**72.22**	**61.76**	**77.78**	**68.85**	**72.73**	**82.76**	**77.42**	**69.54**	**84.68**	**76.36**

TABLE 3: Résultats de l'évaluation extrinsèque (anglais, chinois, grec, polonais, russe et corpus complet) avec N/A représentant les valeurs non-calculables (nombre nul de vrais positifs).

La Table 4 récapitule pour chaque langue et chaque mesure, l'outil ou le chaînage le plus performant et le score associé. BP est souvent l'outil le plus efficace (première place dans un cas sur deux). Toutefois, il existe de nombreuses configurations où d'autres outils ou chaînages (notamment BP–JTA et BP–JTS) offrent de meilleurs résultats.

Si BP donne les meilleurs résultats sur l'évaluation intrinsèque, le choix de ce détoureur semble moins évident lorsque l'on s'intéresse à la tâche. De plus, la qualité d'extraction du contenu informatif est fortement dépendante de la langue traitée. Le corpus polonais est le seul pour lequel on observe des résultats comparables sur les deux évaluations ($F_1 = 72.75$ pour l'évaluation intrinsèque TM, $F_1 = 72.13$ pour l'évaluation extrinsèque). Le contenu est correctement extrait et les résultats de l'évaluation extrinsèque sont très bons et même meilleurs dans certains cas que ceux obtenus sur la version détourée manuellement. Ceci semble dû au fait que les résultats pour le polonais du système DANIEL sont les plus faibles des cinq langues étudiées ($F_1 = 68.85$), la segmentation opérée par le système est plus sujette à caution sur ce sous-corpus. Les résultats obtenus sur le russe sont assez différents de ceux obtenus sur le polonais, bien que les deux langues soient apparentées. L'évaluation intrinsèque et l'évaluation extrinsèque donnent de mauvais résultats. Par ailleurs, le russe est la seule langue pour laquelle BP ne présente pas les meilleures performances pour aucun des traits évalués.

BP donne les meilleurs résultats sur le chinois, et dans une moindre mesure sur le grec et le polonais. Dès lors que

8. Les résultats sont différents de ceux présentés par Brixtel *et al.* (2013) car seuls les documents dont les sources HTML ont été retrouvées sont comptabilisés.

	TO			CAR		
	P	R	F_1	P	R	F_1
anglais	BP-JTA (86.98)	BP (92.02)	BP (88.89)	BP-JTA (87.60)	BP (91.03)	BP (87.87)
chinois	BP (61.32)	BP (52.90)	BP (56.80)	BP (77.12)	BP (63.55)	BP (69.68)
grec	BP-JTA (93.62)	BP (96.48)	BP (94.10)	BP (87.58)	BP (91.59)	BP (89.54)
polonais	BP-JTA (85.24)	JTS-BP (87.54)	BP (84.26)	BP-JTA (82.95)	JTS-BP (82.50)	BP (81.42)
russe	BP-JTA (67.77)	JTS-BP (86.81)	BP-JTA (69.92)	BP-JTA (53.65)	JTS-BP (79.96)	JTA-BP (59.56)
toutes	NC (98.53)	BP (88.89)	BP (85.20)	NC (96.65)	BP (81.12)	BP (78.97)

	TM			Tâche		
	P	R	F_1	P	R	F_1
anglais	BP-JTA (81.09)	BP (93.59)	BP-JTS (79.79)	BP-JTS (65.22)	NC (60.00)	NC (59.15)
chinois	JT-BP (89.91)	BP (67.99)	BP (75.34)	NCT (83.33)	BP (93.75)	BP (85.71)
grec	BP-JTA (86.18)	BP (91.67)	BP-JTS (82.90)	BP (85.71)	JTA (76.47)	BP-JTA (68.57)
polonais	BP-JTA (76.27)	BP (85.57)	BP (72.75)	NCT (82.35)	BP-JTS (81.48)	BP-JTS (72.13)
russe	BP-JTA (48.63)	JTS-BP (86.68)	JTA-BP (58.74)	NCNL (100)	BP-JT, JTA (79.31)	JTA (80.70)
toutes	NC (89.01)	BP (85.42)	BP (73.48)	BP (74.68)	BP-JT (62.10)	BP-JTA (64.98)

TABLE 4: Outils et chaînages offrant les meilleurs scores pour chaque métrique d'évaluation et sous-corpus.

l'on intègre le balisage dans l'évaluation (TM), l'écart entre BP et JT se réduit, exception faite du chinois. NC affiche des résultats variables (en langue et en type d'évaluation) pour être considéré comme fiable. Les meilleurs résultats sur l'évaluation extrinsèque sont obtenus par le chaînage BP-JTA : BP détecte plus de contenu que nécessaire, contenu qui est ensuite rogné par JTA. Les résultats sur l'évaluation intrinsèque TM sont les plus corrélés avec ceux de l'évaluation par la tâche. L'évaluation TM apporte donc le plus d'indices sur le choix du détoureur. Ceci est dû au fait que DANIEL, l'outil utilisé pour l'évaluation extrinsèque, requiert des informations de contenu et de structure.

5 Discussion

Nous avons procédé à une évaluation intrinsèque et extrinsèque d'outils de détourage. Notre questionnement a été le suivant : quelle influence a le détourage sur les résultats d'un système placé en aval de la chaîne de traitement. Nous avons montré que les détoureurs offrant les meilleurs résultats dans l'évaluation intrinsèque ne garantissent nullement de bons résultats pour l'évaluation extrinsèque. Enfin, les résultats ne sont pas constants entre les langues. De manière générale, nous voyons le détourage comme l'illustration d'une problématique rarement abordée qui est celle des erreurs en cascade dans une chaîne de traitement de TAL. L'évaluation de l'influence d'un outil sur les résultats d'un autre n'est pas suffisamment explorée, surtout dans des processus impliquant le chaînage de nombreux outils (détourage, lemmatisation, étiquetage . . .). Le choix d'un outil de détourage ne devrait se faire qu'en fonction de la tâche visée car les résultats obtenus sur l'évaluation intrinsèque ne donnent que de maigres indications sur la pertinence de l'outil en conditions réelles.

Pour une analyse plus globale des résultats présentés dans cet article, nous pouvons reprendre les termes utilisés par les organisateurs de CLEANEVAL. Le détourage est une tâche peu gratifiante et il est sans doute aussi peu gratifiant d'évaluer un système de traitement automatique conjointement avec les modules de nettoyage dont il dépend. Cela conduit à présenter des résultats détériorés, parfois de manière très significative, ce qui pourrait inciter à ne présenter que des résultats obtenus sur des corpus « idéaux ». Il nous semble au contraire tout à fait justifié d'évaluer l'intégralité du processus de traitement en plus des différentes étapes qui le constituent. Évaluer comment les *conditions de laboratoire*, des textes parfaitement détourés, influent sur les résultats devrait alors être un souci plus constant des travaux de TAL. S'il n'existe pas de méthode permettant de détourer efficacement les pages Web autrement que *ad hoc* à un site, alors il convient d'admettre que le détourage n'appartient pas au domaine de l'ingénierie mais reste un champ de recherche que le TAL devrait investiguer.

Références

BARONI M., CHANTREE F., KILGARRIFF A. & SHAROFF S. (2008). Cleaneval : a competition for cleaning web pages. In *Actes du 4ème Workshop Web as Corpus, LREC 2008* : European Language Resources Association.

BRIXTEL R., LEJEUNE G., DOUCET A. & LUCAS N. (2013). Any Language Early Detection of Epidemic Diseases from Web News Streams. In *International Conference on Healthcare Informatics (ICHI)*, p. 159–168.

CHAKRABARTI D., KUMAR R. & PUNERA K. (2008). A graph-theoretic approach to webpage segmentation. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, p. 377–386, New York, NY, USA : ACM.

DAS S. N., VIJAYARAGHAVAN P. K. & MATHEW M. (2012). Article : Eliminating noisy information in web pages using featured dom tree. *International Journal of Applied Information Systems*, **2**(2), 27–34. Published by Foundation of Computer Science, New York, USA.

ENDRÉDY I. & NOVÁK A. (2013). More effective boilerplate removal – the goldminer algorithm. *Polibits*, **48**, 79–83.

EVERT S. (2008). A lightweight and efficient tool for cleaning web pages. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*.

FERRARESI A., ZANCHETTA E., BARONI M. & BERNARDINI S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*.

KOHLSCHÜTTER C., FANKHAUSER P. & NEJDL W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, p. 441–450, New York, NY, USA : ACM.

PASTERNACK J. & ROTH D. (2009). Extracting article text from the web with maximum subsequence segmentation. In *WWW*, p. 971–980.

POMIKÁLEK J. (2011). Removing boilerplate and duplicate content from web corpora. *Disertacnı práce, Masarykova univerzita, Fakulta informatiky*.

RATCLIFF J. W. & METZENER D. E. (1988). Pattern matching : The gestalt approach. *Dr. Dobbs Journal*, **13**(7), 46, 47, 59–51, 68–72.

SPOUSTA M., MAREK M. & PECINA P. (2008). Victor : the Web-Page Cleaning Tool. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*.

VIEIRA K., DA SILVA A. S., PINTO N., DE MOURA E. S., CAVALCANTI J. A. M. B. & FREIRE J. (2006). A fast and robust method for web page template detection and removal. In *ACM international conference on Information and knowledge management*, CIKM '06, p. 258–267, New York, NY, USA : ACM.